

早稲田大学大学院情報生産システム研究科

博士論文審査結果報告書

論 文 題 目

**Study on SVM Classifiers for Imbalanced Data
Classification Using Quasi-Linear Kernel**

申 請 者

Peifeng LIANG

情報生産システム工学専攻
ニューロコンピューティング研究

2021 年 1 月

パターン分類は機械学習において重要な研究課題の一つであるが、現実世界の分類問題では、不均衡データを扱わなければならない場合が多い。訓練データが不均衡な場合に、特段の工夫をせずに分類モデルを構成すると少数派クラスの分類精度が低いモデルになるという問題点がある。一方、サポートベクターマシン (Support Vector Machine: SVM) は、マージン最大化によって優れた汎化性を持つ線形分類器として近年多くの注目を集めている。また、SVM はカーネル法によりデータを高次元特徴空間に写像することによって非線形分類問題を扱うことを可能にしている。従来の SVM モデルでは基本カーネルあるいは基本カーネルの組合せで合成されるカーネル関数を利用しているが、基本カーネル関数の形が固定されているので、分離境界線に関する先見情報があっても利用しにくいという問題がある。それに対して先行研究 (Zhou et al., 2014, Li et al., 2016) では、分離境界線を近似する Coarse-to-Fine という Two-step モデリング法が提案されている。そこでは、Two-step モデリング法の Coarse-step で、回帰ベクターの内積で定義される中間モデルを構築し、Fine-step で、Coarse-step で得られた中間モデルを準線形カーネル関数として利用し、マージン最大化で SVM 分類器を構築する。このように、データが均衡である場合において、体系的なカーネル関数の合成法は確立されており、その有効性も確認されているが、不均衡データに対応できないという問題点を残している。

そこで本研究では、準線形カーネルという中間モデルの構築を活用して、3 つのシナリオにおいて、不均衡データに対応できる準線形 SVM 分類器の構築を行っている。すなわち、1) クラス間のみでなくクラス内にもデータ不均衡がある場合に対して、準線形 SVM モデルに埋め込まれている複数のローカル線形モデルを活用してローカルオフセット調整機能を有する準線形 SVM 分類器を構築している (第 2 章)。2) クラス間にデータ不均衡があるだけでなく少数派クラスのデータが非常に少ない場合に対して、準線形カーネルに基づいた特徴空間に SMOTE (Synthetic Minority Oversampling Technique) 法を実装する方法を提案し少数派クラスのサンプルがカーネル空間で合成される準線形 SVM 分類器を構築している (第 3 章)。3) 多数派クラスのデータしか学習に利用できない One-Class 分類の場合に対して、ゲート付き線形ネットワークを導入し特徴空間において区分線形モデルを構築することによって、高性能な One-Class 準線形 SVM 分類器を構築している (第 4 章)。

以下に、本論文の構成と各章の概略について述べ、評価を与える。

第 1 章では、不均衡データ分類、SVM 分類器およびカーネル法、準線形カーネルについて述べ、本研究の背景・目的を明らかにしている。

第 2 章では、クラス間のみでなくクラス内のデータ不均衡に対応できるローカルオフセット調整機能を有する準線形 SVM 分類器の構築について述べている。従来の重み付き SVM 分類器やオフセット調整可能な SVM 分類器はクラス間のデータ不均衡にしか対応できず、クラス内のデータ不均衡に対応できない

という問題点がある。そこで本論文では、まず、データのラベル情報を活用した不均衡データに対応できるクラスタリング法を提案し、分離境界線に沿ってパーティショニングを行う。次に、パーティショニング情報を利用して準線形カーネル関数を合成し準線形 SVM 分類器を構築する。さらに、準線形 SVM 分類器に埋め込まれている複数の局所的線形モデルをそれぞれのパーティションにおける F-score 最大化線形 SVM 分類器との比較をすることにより、データ不均衡によるローカルオフセットを求める。これに基づいて、クラス内のデータ不均衡にも対応できるローカルオフセット調整機能を有する準線形 SVM 分類器を実現している。性能評価のための数値例では、提案された準線形 SVM と従来の重み付き SVM (WSVM, Y.Huang et al. 2005)、オフセット付き SVM (WHM-SVM, B.Li et al. 2008) を実世界の 14 のベンチマークデータセットに適用して比較し、総合評価 F-score では、13 のデータセットにおいて提案された準線形 SVM が従来の WSVM と WHM-SVM より優れていることが示され、残りの 1 データセットのみ従来の WHM-SVM が優れていた。よって、提案されたクラス内データ不均衡に対応した準線形 SVM 分類器の有効性が明らかとなっている。

第 3 章では、クラス間に不均衡があるだけでなく少数派クラスのデータが非常に少ない場合において、オーバーサンプリング機能付き準線形 SVM 分類器の構築について述べている。少数派クラスのサンプルが非常に少ないデータ不均衡問題に対処するため、有効なオーバーサンプリング法として知られている SMOTE では、少数派クラスのサンプルを新しく線形合成で生成し、バランスが取れたデータセットを提供することで不均衡問題を解決している。しかし、従来の SMOTE では、クラス間の分離境界線が非線形である場合に、多数派クラスと重複した誤った少数派クラスのサンプルを生成して分類が難しくなる可能性がある。そこで、本論文では、準線形カーネルに基づいた多線形特徴空間 (Multi-Linear Feature Space: MLFS) を導入し、SMOTE が MLFS に実装される方法を提案している。さらに、従来とは異なり少数派クラスのサンプルがカーネル空間で生成される方法を導入し、準線形 SVM 分類器の構築を行っている。性能評価のための数値例では、提案された準線形 SVM 分類器と従来の State of the Art オーバーサンプリング機能付き SVM 分類器: BSMOTE (H.Han et al. 2011)、MSMOTE (S.Barua et al. 2014)、SSMOTE (B.Zhou et al. 2017) を実世界の 18 のベンチマークデータセットに適用して比較した結果、総合評価 F-score では、16 のデータセットにおいて提案された準線形 SVM 分類器が 3 つの従来法より優れており、BSMOTE と MSMOTE SVM 分類器がそれぞれ残りの 1 データセットで優れていた。特に、real-sim という高次元(20958 次元)を持つデータセットに対して、従来法の F-score はそれぞれ 0.526、0.595 と 0.599 であるのに対して、提案法は 0.957 であり、提案された準線形 SVM 分類器の有効性が明らかとなっている。

第 4 章では、多数派クラスのデータしか学習に利用できない **One-Class** 準線形 **SVM** 分類器の構築について述べている。**One-Class** 分類の場合には、少数派クラスのデータが少なすぎて学習に利用できない状況を想定しているため、分類器の構築がより困難となっている。また、**One-Class** 分類では、多数派クラスのデータ（正常サンプル）に基づいて構築された分類器は少数派クラスのデータ（異常サンプル）の検出も目的としている。異常サンプル検出の正確度を向上するため、正常サンプルを囲む可能な限りコンパクトな分離境界線が要求される。そこで本論文では、ゲート付線形ネットワークの導入および 2 進数配列となるゲート制御信号を **Winner-Take-All Autoencoder** で生成することにより、アクセスしにくい高次元特徴空間において柔軟な区分線形モデルの構築を実現している。これによって、**One-Class SVM** 用準線形カーネル関数を合成し、高性能な **One-Class** 準線形 **SVM** 分類器の構築を行っている。性能評価のための数値例では、提案された **One-Class** 準線形 **SVM** 分類器と従来の **One-Class SVM** 分類器を実世界の 14 のベンチマークデータセットに適用して比較し、総合評価 **F-score** においては 13 のデータセットで、異常サンプル検出の正確度においては 14 のすべてのデータセットで、提案法が従来法より優れていた。**F-score** と正確度の平均値では、従来法はそれぞれ 0.808 と 0.870 であるのに対して、提案法は 0.853 と 0.910 となり、提案された **One-Class** 準線形 **SVM** の有効性が明らかとなっている。

第 5 章では結論として、本研究の成果について総括し、今後の研究課題を論じている。

以上を要約すると、本論文では、準線形カーネルという中間モデルの構築を活用して、異なるシナリオの不均衡データに対応できる 3 つの **SVM** 分類器：1) ローカルオフセット調整機能を有する準線形 **SVM** 分類器、2) 特徴空間に実装される **SMOTE** オーバーサンプリング機能付き準線形 **SVM** 分類器、3) 特徴空間において構築される区分線形モデルに基づいた **One-Class** 準線形 **SVM** 分類器の提案を行っている。また、提案された準線形 **SVM** 分類器を複数の実世界のデータセットに適用しその有効性を明らかにしている。これらの成果は機械学習、ニューロコンピューティング分野の発展に寄与するところ大である。よって、本論文は、博士（工学）の学位論文として価値あるものと認める。

2020 年 12 月 16 日

審査員

主査	早稲田大学	教授	博士(情報工学)(九州工業大学)	古月 敬之
副査	早稲田大学	教授	工学博士(早稲田大学)	吉江 修
	早稲田大学	教授	博士(工学)(早稲田大学)	藤村 茂